Web Browsing Behaviors Based Age Detection

Misha Kakkar, Divya Upadhyay

Abstract— Users basic attributes like age, gender location etc… plays an essential role in today's web applications. Previous research shows that there is relationship between users' browsing behavior and their basic characteristics. In this paper we made an approach to detect a user's age depending on his web browsing history. The user's web browsing behaviors is treated as a variable to propagate age information between different users. Artificial neural network tool is used for this purpose. Uses are divided into two different categories of adult and youngsters. The result is 93.7% accurate.

Index Terms— Age prediction, Browsing behavior, Artificial.

I. INTRODUCTION

Several common network services, such as searching, websites etc, begin to give more attentions to modified service for an improved user experience. My Yahoo! and Google Personal are excellent exemplar among these approaches. My Yahoo! permit users to build their choice explicitly and only show sectors and information which they may be concerned in [6]. Google Personal arranges users' search outcome according to their search records including their earlier search results and news heading clicked [1]. Along with the success of common web service, online promotion is increasing rapidly in current years, in which behavioral targeting is becoming popular. Behavior targeting facilitates promoters to aim right users upon their behaviors while surfing online. Also there are some website contents are not appropriate for children below 18 and it is not an easy task to stop children from surfing these websites content. Still, demographic information is generally not simple to find. Internet users are hesitant to expose this sort of private information to public. Another method to guess users' demographic information is then of large interest to both commerce and academic circles. In Koppel's work [8], writing styles of the bloggers are used to forecast their actual demographic information. However, very small figure of network users writes blogs [2]. In contrast, the bulk of users surf news, goods, or other web pages through internet, which gives us a huge amount of web-page click-through log data. Earlier learning illustrate that there is connection between users' browsing behavior and their demographic feature. As reported in "Computerworld" [5], 74% of adult search for health or medical information online, 34% of adult search for religious information from the internet. Similar observation happens in movie field, where demographic information shows a relationship to the sort of the movies the viewers appreciate. Action and love for grown person, or cartoon for teens are general link between movie type and viewer's demographic type. So the variety of the user's online browsing behavior can be used to decide an unidentified user's demographic characteristic such as age.

In this paper we examine the difficulty of guessing the website users' age based on their browsing actions, in which the type of website viewed is treated as a concealed variable to predict the age of different users. The solution consists of information taken from two different categories of users adult and youngsters which is analyzed to predict the user's age based on their profile and browsing behavior and then, a supervised neural network model is trained to predict a webpage user's age i.e. the probability distribution of the ages of a given Webpage's readers. Based on the error analysis, the prediction model resulted from the above steps gives a good accuracy.

The remaining paper is arranged in seven different sections. In Section2, related work is presented. In Section 3, the prediction problem is stated. In Section 4, the solution is proposed for the same using artificial neural network. The experimental results are shown and compared in Section 5. Then conclusion is drawn and highlighted future research directions in Section 6 and section 7 respectively.

II. RELATED WORK

In this section we in brief present some of the study literature associated to age prediction. Nguzen, Smith and Rose frames [4] age prediction of the text writer using a regression model. Data set is created using blog data, telephone conversations and online forum posts. Domain adaptation technique is used to train a model which combines data from all the sources and also work on every source separately. The model gets a correlation up to 0.74 and means absolute error between 4.1 to 6.8 years.

Previous study on demographic prediction generally paid attention on modeling the variety of the linguistics writing and speaking styles related with the demographic characteristic that also mainly with gender of the user .Koppel [8] analyzed that there are major variations in both writing style and content between authors of different ages. Based on these differences on blog's content and style, Multi-Class Real Winnow algorithm was used by them to learning models that categorize blogs [1] according to the writer gender and age, and 76.2% accurateness on age groups in 13-17, 23-27, and 33-42. Hu, Zeng, Niu and Chen [6], investigated the difficulty of predicting internet users' demographic attributes (gender and age) based on their browsing history, in which the viewed webpage information is treated as a concealed variable to propagate age and gender between different users. Their research outcome on a real large page click-through log indicates that the proposed algorithm had achieved 60.3% accuracy.

Manuscript received on March, 2013.

Misha Kakkar, Computer Science & Engg., ASET-Amity University, Noida, India.

Divya Upadhyay, Computer Science & Engg., ASET-Amity University, Noida, India.

*Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication*

These studies mainly focus on classifying users' age based on browsing habit. As far as we know, there is a confusing work on calculating users' age according to what they browsed on the internet.

A. *Review Stage*

Submit your manuscript electronically for review.

B. *Final Stage*

When you submit your final version, after your paper has been accepted, prepare it in two-column format, including figures and tables.

C. *Figures*

As said, to insert images in *Word*, position the cursor at the insertion point and either use Insert | Picture | From File or copy the image to the Windows clipboard and then Edit | Paste Special | Picture (with "Float over text" unchecked).

The authors of the accepted manuscripts will be given a copyright form and the form should accompany your final submission.

III. PROBLEM DEFINITION

Prior to the introduction to technology for the prediction of user age, we state the difficulty in this section.User's attributes i.e. age is presented as a vector age. The age prediction classifies users into two categories:

Table 1: Age Group

Group	Age
Youngster	<18
Adult	=>18

The browsing history is defined as a record set having 16 columns which contains the probability of different web pages viewed by the corresponding user. This data is collected by conducting a survey among internet user of age groups between 12 to 50 years The common web page behavior of inter usage that can be used to predict age includes email usage, social networking sites, job search, pc to mobile usage, downloading and other…

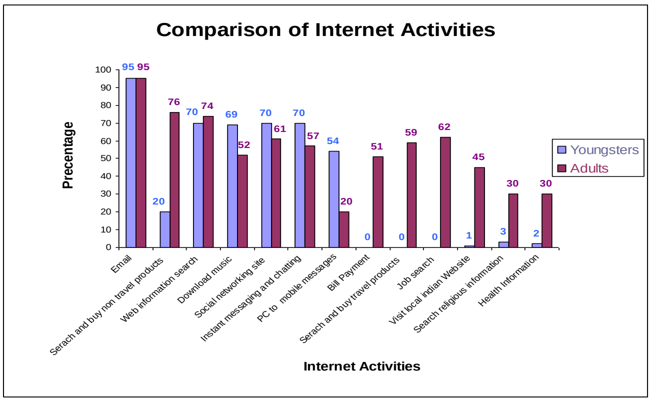


Figure I: Comparison of Internet Activities

Figure I shows the comparison of internet activities among various users of both the categories. The proposed methodology focuses on developing a system for behavioral analysis which can predict internet user age automatically with computer aid and with any human intervention.

Artificial Neural Network is a powerful classification model, very useful where the underlining factors may display chaotic associations. Due to this the proposed methodology uses ANN to solve the problem.

IV. AGE PREDICTION

ANN design includes the selection of ANN model, architecture and learning algorithm based on the need of the problem at hand.

A. *ANN model Selection*

A multilayer perceptron (MLP) with back propagation algorithm having one input layer, hidden layer(s) and one output layer can be used as dominant tool for data analysis [8]. The numbers of hidden layers are chosen not only to reduce the network complexity but also to increase its computational power.

B. *Input layer*

Comscore media survey [3] was studied and a questionnaire was prepared which was then circulated among people of various age groups for assessment. Results of analysis were shown in figure1 based on which 16 input factors were decided as input parameters.

C. *Hidden Layer*

Hidden layer(s) gives the network its generalizing ability. There is no fix law for formulating the number of hidden layers and hidden modes in a network. A rough estimation can be achieved by the geometric pyramid rule projected by masters [9]. Various ANN's having different configuration was created and there accuracy was compared in figure 2.

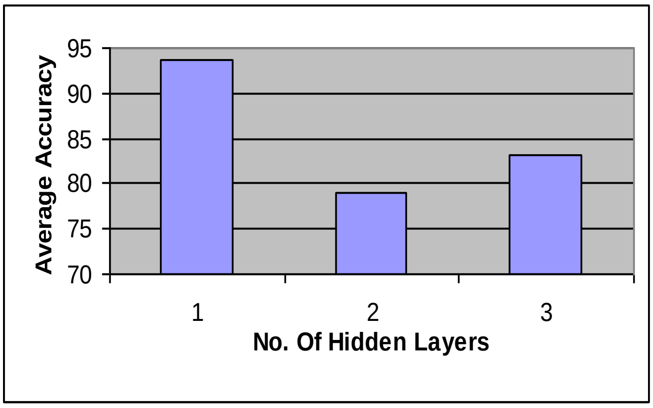


Figure II: Hidden Layer Comparison

D. *Output Layer*

The number of nodes in the output layer is decided according to the application's output. Since the neural network is used to predict the age group of the user the output node is one.

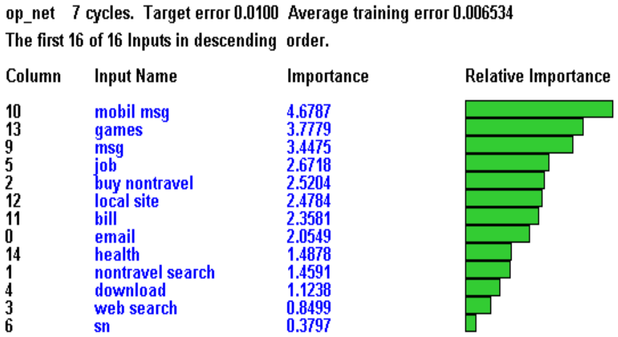
E. *Neural Network Training*

Easy NN tool was used to build and trained the neural network. Data set was created before training the network. There were in all 250 records used to train the network. MLP was trained in supervised mode using back propagation algorithm in which target error were set as 0.01. Learning weight and momentum was set as 0.4 and 0.6 respectively.

At the time of training the MLP automatically learns the association between input and output because of which in its application phase when those similar patterns are applied, it produces required results.

V. RESULTS

As discussed earlier there is no fixed rule to evaluate the number of hidden layers in neural network. We created multiple networks having varied hidden layers. To evaluate the accuracy of network prediction testing data set having 150 records was applied and it was found that MLP having only one hidden layer gives maximum accuracy (Figure II). The most accurate MLP has 16 nodes in its input layer, 8 nodes in its hidden layers and one node in its output layer. It gives maximum importance to factors like PC-Mobile messages, gaming, instant messages and chatting, job search and online shopping.



```
op_net   7 cycles.  Target error 0.0100  Average training error 0.006534
The first 16 of 16 Inputs in descending  order.

Column   Input Name          Importance     Relative Importance

10       mobil msg           4.6787
13       games               3.7779
9        msg                 3.4475
5        job                 2.6718
2        buy nontravel       2.5204
12       local site          2.4784
11       bill                2.3581
0        email               2.0549
14       health              1.4878
1        nontravel search    1.4591
4        download            1.1238
3        web search          0.8499
6        sn                  0.3797
```

Figure III: Relative Importance of the Factors

VI. CONCLUSION

Through this paper, a method has been proposed for predicting the internet user's age based on their browsing history using Artificial Neural Network. It establishes suitability of non-linear ANN as predictive tool for internet user population. Final result gives an accuracy of 93.7%. If implemented as a web service, this tool can be used to stop youngsters from accessing censored contents online.

FUTURE WORK

There are huge possibilities for future study in this area. Till now, the age group is divided into two categories; we are planning to further narrow down the age group.

We plan to extend our work on other attributes as gender, location, occupation etc.

REFRENCES

1. A. Lenhart, S. Fox. Bloggers: A portrait of the internet's new storytellers. http://www.pewinternet.org/pdfs/PIP%20Bloggers%20Report%20July%2019%202006.pdf
2. Burger, J. and Henderson, J. (2006), An Exploration of Observable Features Related to Blogger Age, in `Computational Approaches to Analyzing Weblogs, AAAI Spring Symposium.
3. ComScore Media matrix, March 2011: India: Digital Market Overview, Digital Strategy Consulting and Digital Training Academy.
4. Dong Nguyen Noah A. Smith Carolyn P. Ros´e: Author Age Prediction from Text using Linear Regression, Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Portland, OR, USA, 24 June 2011, pp. 115–123.
5. Eric Auchard : 2005, Study: Men want facts, women seek personal connections on Web: Computerworld magazine.
6. Hu, J., Zeng, H. J., Li, H., Niu, C. and Chen, Z. (2007), Demographic prediction based on user's browsing behavior, in `WWW '07: Proceedings of the 16th international conference on World Wide Web', ACM, pp. 151-160.
7. Hassoun Md. H.: 1995, Fundamentals of Neural Network MIT, Liberty of Congress
8. M. Koppel, J. Schler, S. Argamon, and J.W. Pennebaker. Effects of age and gender on blogging. In AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs, 2006.
9. Masters, T.: 1993, Practical neural network recipes in C++. San Diego, CA, USA: Academic Press Professional, Inc.
10. Murray, D. and Durrell, K. (2000), Inferring Demographic Attributes of Anonymous Internet Users, in `WEBKDD '99: Revised Papers from the International Workshop on Web Usage Analysis and User Profiling', pp. 7-20

AUTHRS PROFILE

Misha Kakkar, received her M.E from Panjab Engineering College, Chandigarh. Presently she is working as assistant professor in Amity University, Noida, Uttar Pradesh, India. Her research area includes Artificial Neural Network and Software Engineering.

Divya Upadhyay, received her M.Tech in Information Security from GGSIP University, New Delhi. Presently she is working as an Assistant Professor in CSE Department, Amity University, Noida, Uttar Pradesh, India. Her Research area includes Information Security and Security Engineering